MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2019/2020

TDM 2151– DATA MINING AND MACHINE LEARNING

(All sections / Groups)

25 OCTOBER 2019
9.00am – 11.00am
(2 Hours)

INSTRUCTIONS TO STUDENT

1. This Question paper consists of 4 pages with 5 Questions only.

2. Answer all **FIVE** questions. All questions carry equal marks and the distribution of the marks for each question is given.

3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

a. Discuss briefly, whether each of the following activities is a data-mining task or not. (4 marks)

   i. Dividing the customers of a company according to their profitability.

   ii. Predicting the future stock price of a company using historical records.

b. Dimensionality reduction is quite important in many data mining projects. Discuss **THREE (3)** importance of dimensionality reduction. (3 marks)

c. Suppose that the data for analysis include the attribute age of 10 children. The age values are 10, 8, 14, 12, 15, 20, 9, 17, 10, 15 Using min-max normalisation, transform the age value of 15 and 8 onto the range [0.0, 1.0]. (3 marks)

QUESTION 2

a. Compare k-Nearest Neighbor Classifiers and Decision Trees. What are the **THREE (3)** main differences between these two approaches? (3 marks)

b. What is overfitting? What are the characteristics of overfitting? What can be done in the context of decision trees to overcome overfitting? (3 marks)

c. A test for a disease gives a correct positive result with a probability of 0.95 when the disease is present, but gives an incorrect positive result (false positive) with a probability of 0.15 when the disease is not present. If 5% of the population has the disease, and Jean tests positive to the test, what is the probability Jean really has the disease? (4 marks)

Continued .....

QUESTION 3

a. Briefly describe the general objective of Association Rule mining.
(2 marks)

b. Consider the following transaction database in **Table 1a** and Association Rules in **Table 1b**

Table 1a: Transaction Database

Tid	Item
T1	bread, butter, milk, sugar, eggs
T2	water, bread, cereal, milk
T3	cheese, butter, eggs, milk, bread
T4	bread, milk
T5	butter, milk, eggs

Table 1b: Association Rules

A. {eggs, cheese}→ {butter}
B. {bread} → {eggs}
C. {milk, butter} → {bread}
D. {water} → {sugar}
E. {cereal} → {milk}

i. Compute the support and confidence of the association rule A to E
(3 marks)

ii. Find strong association rules among the association rule A to E when minimum support threshold is 40% and minimum confidence threshold is 60%.
(1 mark)

c. Contingency **Table 2** summarises a supermarket transaction data, where it contains Bread and Milk.

Table 2: Contingency Table

	Milk	Not Milk
Bread	3000	1000
Not Bread	800	1200

Based on the given data, is the purchase of Bread independent of the purchase of Milk? If not, what kind of correlation relationship exists between the two?
(4 marks)

Continued .....

QUESTION 4

a. What is the goal of hierarchical clustering? How is it different from partitioning clustering algorithms, such as K-Means? (3 marks)

b. Given the dataset in **Table 3**

Table 3: Dataset

	A	B	C	D	E	F
Axis-X	1	0	2	3	4	3
Axis-Y	0	1	1	0	0	2

Initially the dataset is clustered into two
C1 = (A, B, C)
C2 = (D, E, F)

i. What is the centroid for each cluster? (2 marks)

ii. What is the total square error from initial cluster distribution? (2 marks)

c. **Table 4** shows the results observed from clustering 6000 data points into 3 clusters: A, B and C. What is the F_1-Score with respect to cluster B? (3 marks)

Table 4: Data Points

		Actual		
		A	B	C
Predicted	A	600	400	200
	B	1000	1200	200
	C	400	400	1600

QUESTION 5

a. Answer True or False. Explain your reasoning. (3 marks)

i. Suppose a dataset has 900 cancer-free images and 100 images from cancer patients. If a trained classifier achieves 85% accuracy on this dataset, it is considered a good classifier.

ii. A classifier that attains 100% accuracy on the training set and 70% accuracy on test set is better than a classifier that attains 70% accuracy on the training set and 75% accuracy on test set.

b. Discuss whether Classifier and Clustering can be used together? Justify you answer with examples. (3 marks)

c. Data mining brings a lot of benefits to businesses, society, and governments. Describe **TWO (2)** benefits of data mining to Telecommunication companies? (4 marks)

End of Page